Estimation of structural mean models with multiple instruments

Tom Palmer\textsuperscript{1}, Paul Clarke\textsuperscript{2}, Frank Windmeijer\textsuperscript{2}

\textsuperscript{1}MRC CAiTE Centre, School of Social and Community Medicine, University of Bristol

\textsuperscript{2}Department of Economics and CMPO, University of Bristol

Royal Statistical Society, 26 May 2011
Aim

Combine two strands of literature:

• Structural mean models [Biostatistics]

• Generalised Method of Moments estimation [Econometrics]

Rationale:

• Concepts such as G-estimation intimidating

• Estimation with multiple instruments

• Straightforward implementation in Stata and R
Outline

• Introduction to example
• Causal parameters & potential outcomes

• Multiplicative SMM
  – What is GMM?
  – Over-identification test
  – Combining multiple instruments
  – Two step GMM
  – Implementation in Stata
  – Local risk ratios
  – MSMM and MGMM

• Logistic SMM
  – Joint estimation

• Summary
Introduction to example

• Copenhagen General Population study
  – N=55,523
• Instruments:
  – \textit{FTO} (rs9939609) chr16, \textit{MC4R} (rs17782313) chr18 genotypes
  – Associated with obesity in GWAS (0.4, 0.2 BMI units). Frayling 2007, Loos 2008
• Exposure:
  – Overweight (body mass index BMI \left[ \frac{\text{weight}}{\text{height}^2} \right] >25)
• Outcome:
  – Hypertension (high blood pressure \left[ \text{SBP}>140\text{mmHg}, \text{or DBP}>90\text{mmHg}, \text{or taking anti-hypertensives} \right])

\textit{FTO}, \textit{MC4R} genotypes \rightarrow \text{Overweight} \rightarrow \text{Hypertension}
<table>
<thead>
<tr>
<th></th>
<th>No Hypertension</th>
<th>Hypertension</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Overweight</td>
<td>10,066 (42%)</td>
<td>13,909 (58%)</td>
<td>23,975</td>
</tr>
<tr>
<td>Overweight</td>
<td>6,906 (22%)</td>
<td>24,642 (78%)</td>
<td>31,548</td>
</tr>
<tr>
<td>Total</td>
<td>16,972 (31%)</td>
<td>38,551 (69%)</td>
<td>55,523</td>
</tr>
</tbody>
</table>

$\chi^2 P < 0.001$

**Risk ratio 1.35 (1.32, 1.37)**

<table>
<thead>
<tr>
<th>FTO</th>
<th>MC4R</th>
<th>Z</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.20</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.15</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.27</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0.21</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0.07</td>
</tr>
</tbody>
</table>

**P = 0.007**

**P < 0.001**

$R^2 = 0.002$
Causal parameters and potential outcomes

• SMMs defined in terms of potential outcomes Hernan & Robins 2006

• $X$: exposure/treatment, $Y$: outcome, $Z$: IV

• $Y(X=1)$ outcome subject would experience if they were given treatment/exposure under intervention
Potential outcomes for an individual

\[ Y(X=1) \quad Y(X=0) \]
Potential outcomes for whole study

Recent discussion of $G$-estimation: Snowden et al., AJE, 2011

Average treatment effect = $E[Y(X=1)] / E[Y(X=0)]$

Causal risk ratio = $E[Y(X=1)] / E[Y(X=0)]$

Causal odds ratio = $\text{odds}[Y(X=1)] / \text{odds}[Y(X=0)]$
What we observe

\[ E[Y(1) | X=1] - E[Y(0) | X=1] \]

SMMs identify effect of treatment of treated
Multiplicative SMM

$Z$ is instrumental variable  $X$ is exposure  $Y$ is outcome

$Y$, $X$ and $Z$ are binary

$$\frac{E[Y|X, Z]}{E[Y(0)|X, Z]} = \exp\{(\theta_0 + \theta_1 Z) X\}$$

$Y(0)$ is the exposure- or treatment-free potential outcome

...so far ... model non-identified: 2 parameters, 1 equation

No effect modification by $Z$ (NEM):  $\theta_1 = 0$

$\theta_0$: log causal risk ratio

Conditional mean independence (CMI) from IV assumptions:

$$E[Y(0)|Z = 1] = E[Y(0)|Z = 0] = E[Y(0)]$$
Moment conditions

$$a_0 = E[Y(0)]$$

Multi-valued instrument/multiple instruments

$$E \left\{ Y \exp(-X\theta_0) - a_0 \right\} | Z = 2 = 0$$
$$E \left\{ Y \exp(-X\theta_0) - a_0 \right\} | Z = 1 = 0$$
$$E \left\{ Y \exp(-X\theta_0) - a_0 \right\} | Z = 0 = 0$$

Over-identified:
3 moment conditions, 2 parameters

Exactly identified:
2 moment conditions, 2 parameters... need GMM

$E[] = 0$ since $Z$ independent of $Y$ given $X$: exclusion restriction

If no $E[Y(0)]$ – need to centre the instruments;
Vansteelandt & Goetghebeur, JRSS B, 2003
What is GMM?

Designed to estimate over-identified models
GMM minimises quadratic form wrt parameters to be estimated

\[
\hat{\delta} = \arg \min_{\delta} \left( \frac{1}{n} \sum_{i=1}^{n} g_i(\delta) \right)' \cdot W_n^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} g_i(\delta) \right)
\]

\[
\begin{align*}
\{ Y \exp(-X\theta_0) - \alpha_0 \} & Z_0 \\
\{ Y \exp(-X\theta_0) - \alpha_0 \} & Z_1 \\
\{ Y \exp(-X\theta_0) - \alpha_0 \} & Z_2
\end{align*}
\]

\[W_n^{-1}\]

\[Z_0 \quad Z_1 \quad Z_2\]

\[Z_0 \quad Z_1 \quad Z_2\]

\[W^{-1}\] affects efficiency not consistency: one step/two step GMM
Over-identification test

Profiling over quadratic form \((Q)\) for a single parameter

- Single instrument – exactly identified: \(\min(Q)=0\)
- Multiple instruments – over identified: \(\min(Q)\) should be close enough to 0 as given by Hansen over-id test statistic, \(Q \sim \chi^2_{m-p}\) when moments valid
- Not rejecting the over-id test \textit{doesn’t} mean the IV assumptions hold

\[\chi^2_{2,0.95}=5.99\]
Combining multiple instruments

How does GMM treat multiple instruments?

The instruments get combined into the projection \( S (S'S)^{-1} S'D \), i.e. a constant 1 and the linear projection of \( \frac{y_i}{\exp(x_i\theta)} x_i \) on \( s_i \), the projection as proposed by Bowden and Vansteelandt (2010).

GMM satisfies

\[
D'S (S'S)^{-1} S'v = 0
\]

\[
D = \{d'_i\} ; \quad S = \{s'_i\} ; \quad v = \{v_i\}
\]

\[
d_i = \left( \begin{array}{c}
1 \\
\frac{y_i}{\exp(x_i\theta)} x_i
\end{array} \right) ; \quad v_i = \frac{y_i}{\exp(x_i\theta)} - \alpha
\]
Two step GMM

Step 1: Estimate parameters and \( W \)
Step 2: repeat optimization starting from step 1 estimate of \( W \)

\[
\hat{\delta}_2 = \arg \min_{\delta} \left( \frac{1}{n} \sum_{i=1}^{n} g_i(\delta) \right)' W_n^{-1} \left( \hat{\delta}_1 \right) \left( \frac{1}{n} \sum_{i=1}^{n} g_i(\delta) \right)
\]

Two-step GMM is efficient because it’s Vcov matrix is the smallest (Chamberlain 1987)

One step: \( \sqrt{n} \left( \hat{\delta}_1 - \delta_0 \right) \xrightarrow{d} N \left( 0, (C_0'WC_0)^{-1} C_0 W \Omega_0 WC_0 (C_0'WC_0)^{-1} \right) \)

Two step: \( \sqrt{n} \left( \hat{\delta}_2 - \delta_0 \right) \xrightarrow{d} N \left( 0, (C_0'\Omega_0 C_0)^{-1} \right) \)
MSMM implementation in Stata

gmm command (Stata version 11)

Moment condition

gmm \((y*exp(-x*\{theta\}) - \{ey0\})\), instruments(z1 z2 z3)

lincom [theta]:_cons, eform

Causal risk ratio

estat overid

Over-identification test
Two step GMM

E[Y(0)] = 0.58 (0.50, 0.65)
Causal risk ratio = 1.36 (1.08, 1.72)
Observational and IV estimate in example

Gamma (log link)

1 1.2 1.4 1.6 1.8

Causal risk ratio (log scale)

1.35 (1.33, 1.36)

1.36 (1.08, 1.72)
Local risk ratios

• Identification depends on NEM ... what happens if it doesn’t hold?
• Alternative assumption of monotonicity: \( X(Z_k) \geq X(Z_{k-1}) \)
• Local Average Treatment Effect (LATE): effect among those whose exposures are changed (upwardly) by changing (counterfactually) the IV from \( Z_{k-1} \) to \( Z_k \)

\[
\alpha_{\text{All}} = \lambda_1 \alpha_{1,0} + \lambda_2 \alpha_{2,1} + \lambda_3 \alpha_{3,2}
\]

Linear IV: Imbens & Angrist 1994

**MSMM:** We show a similar result holds for MSMM (\( X, Y \): binary)

\[
e_z^\theta = \sum_{k=1}^{K} \tau_k e_{k,k-1}^\theta
\]

...weighted average of risk ratios

... rather than log risk ratios!
Local risk ratios in the example

Check: \((0.10 \times 2.21) + (0.81 \times 1.11) + (0.09 \times 2.6)\)
MSMM and MGMM

MGMM: Mullahy 1997 – exponential mean model with multiplicative residual

Additive residual: \[ Y = \exp(X\theta) + U \]
\[ E[Z(Y - \exp(X\theta))] = 0 \]

Poisson regression

Multiplicative residual: \[ Y = \exp(X\theta + U) \]
\[ E \left[ \frac{Y - \exp(\alpha_0^* + X\theta_0)}{\exp(\alpha_0^* + X\theta_0)} \mid S \right] = 0 \]
\[ S = (1, Z_1, Z_2)' \]

Proof MSMM = MGMM

Clarke & Windmeijer 2010; Didelez, et al. 2010; Palmer et al., AJE, 2011
MGMM (one step GMM): ivpois for Stata (Nichols 2007)
Logistic SMM

- Implement joint estimation approach within GMM framework
- Vansteelandt & Goetghebeur (2003), Vansteelandt & Bowden (2010)

<table>
<thead>
<tr>
<th>Two-stage estimation</th>
<th>Joint estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stage 1</strong></td>
<td></td>
</tr>
<tr>
<td>Association model:</td>
<td></td>
</tr>
<tr>
<td>predict $Y$ given $X$, $Z$</td>
<td>Estimate association model and causal model together</td>
</tr>
<tr>
<td><strong>Stage 2</strong></td>
<td></td>
</tr>
<tr>
<td>Causal model</td>
<td></td>
</tr>
<tr>
<td>(MSMM/ASMM causal model only)</td>
<td></td>
</tr>
<tr>
<td>Need to correct SEs somehow</td>
<td>SEs automatically correct</td>
</tr>
</tbody>
</table>

Gourieux 1996, Tan 2010
LSMM implementation in Stata

**Two step estimation**

```
logit y x z1 z2 xz1 xz2
matrix from = e(b)
predict xblog, xb
```

Association model: predict \( Y \) given \( X, Z \)

```
gmm (invlogit(xblog - x*{psi}) - {ey0}), instruments(z1 z2)
matrix from = (from,e(b))
```

Causal model – incorrect SEs!

**Joint estimation** – correct SEs!

```
gmm (y - invlogit({logit:x z1 z2 xz1 xz2} + {logitconst}))
   (invlogit({logit:} + {logitconst} - x*{psi}) - {ey0}), ///
   instruments(1:x z1 z2 xz1 xz2) instruments(2:z1 z2) ///
   winitial(unadjusted, independent) from(from)
```

```
lincom [psi]_cons, eform // causal odds ratio
estat overid
```
LSMM Stata output

```
. logit hyp overw Iz1 Iz2 Iz3 Iz1Xoverw Iz2Xoverw Iz3Xoverw

Iteration 0:  log likelihood =  -34179.76
Iteration 1:  log likelihood =  -32895.818
Iteration 2:  log likelihood =  -32885.846
Iteration 3:  log likelihood =  -32885.845

Logistic regression                     Number of obs   =      55523
LR chi2(7)  =    2587.83
Prob > chi2 =     0.0000
Pseudo R2   =     0.0379
Log likelihood =  -32885.845

Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------- ------------- ------------- -------- ------------------
    hyp      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
    overw     .9034696   .0419769    21.52   0.000     .8211964    .9857428
     Iz1     .0023852   .0346439     0.07   0.945    -.0655155    .0702864
     Iz2    -.031613   .0375747    -0.84   0.400    -.105258    .0420328
     Iz3     .0285799   .0598671     0.48   0.633    -.0887574    .1459173
     Iz1Xoverw  .0500117   .0509504     0.98   0.326    -.0498493    .1498727
     Iz2Xoverw  .06952    .0543206     1.28   0.201    -.0369465    .1759864
     Iz3Xoverw  .041216    .0837708     0.49   0.623    -.1229717    .2054037
   _cons     .3295621   .0285043    11.56   0.000     .2736947    .3854295

. matrix from = e(b)
. predict xblog, xb
```

Association model

predicted values of outcome (on logit scale here)
Causal model

Incorrect SEs
Joint estimation

Corrected SEs: causal model SEs ×10
Causal odds ratio = 2.87 (1.25, 6.55)

| exp(b)   | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|---------|-----------|-------|-------|----------------------|
| (1)     | 2.86555   | 1.208417 | 2.50 | 0.013    | 1.253868 – 6.548836 |

Degrees of freedom:
AM: exactly identified
CM: 4 moments – 2 pars
Observational and IV estimate in example

LSMM Logistic

Causal odds ratio (log scale)

2.87 (1.25, 6.55)

2.58 (2.49, 2.68)
Summary

• Estimate SMMs within GMM framework
• GMM optimal combination of multiple instruments
• Two-step GMM is efficient
• Joint estimation for LSMM
• Hansen over-identification test
  – Joint validity of multiple instruments
  – Can help detect violations in NEM & CMI
• Straightforward implementation in Stata and R
References

Bowden & Vansteelandt, Mendelian randomization analysis of case-control data using structural mean models, Stats Med, 2011
Chamberlain, Asymptotic efficiency in estimation with conditional moment restrictions, J Econ, 1987
Clarke & Windmeijer, Identification of causal effects on binary outcomes using structural mean models, Biostatistics, 2010
Didelez et al., Assumptions of IV Methods for Observational Epidemiology, Stat Sci, 2010
Frayling et al., A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity, Science, 2007
Gourieroux et al., Two-stage generalized moment method with applications to regressions with heteroscedasticity of unknown form, J Stat Plan Inf, 1996
Loos et al., Common variants near MC4R are associated with fat mass, weight and risk of obesity, Nature Genetics, 2008
Palmer et al., Instrumental Variable Estimation of Causal Risk Ratios and Causal Odds Ratios in Mendelian Randomization Analyses, AJE, 2011, in press
Snowden et al., Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique, AJE, 2011
Tan, Marginal and Nested Structural Models Using Instrumental Variables, JASA, 2010
Windmeijer & Santos Silva, Endogeneity in Count Data Models: An Application to Demand for Health Care, J Appl Econ, 1997
Windmeijer, GMM for Panel Count Data Models, CEMMAP Working Paper CWP21/06, 2006
Vansteelandt & Goetghebeur, Causal Inference with Generalized Structural Mean Models, JRSS B, 2003
Acknowledgements

• MRC Collaborative grant G0601625
• MRC CAiTE Centre grant G0600705
• ESRC grant RES-060-23-0011
• With thanks to Nuala Sheehan, Vanessa Didelez, Debbie Lawlor, Jonathan Sterne, George Davey Smith, Roger Harbord, Sha Meng, Nic Timpson, Borge Nordestgaard, John Thompson, Martin Tobin.